

**This Page Is Inserted by IFW Operations  
and is not a part of the Official Record**

## **BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- **BLACK BORDERS**
- **TEXT CUT OFF AT TOP, BOTTOM OR SIDES**
- **FADED TEXT**
- **ILLEGIBLE TEXT**
- **SKEWED/SLANTED IMAGES**
- **COLORED PHOTOS**
- **BLACK OR VERY BLACK AND WHITE DARK PHOTOS**
- **GRAY SCALE DOCUMENTS**

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning documents *will not* correct images,  
please do not report the images to the  
Image Problems Mailbox.**

**THIS PAGE BLANK (USPTO)**



## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification: <b>H04J 3/26</b>	A1	(11) International Publication Number: <b>WO 00/04665</b> (43) International Publication Date: <b>27 January 2000 (27.01.2000)</b>
(21) International Application Number: <b>PCT/US99/15071</b>  (22) International Filing Date: <b>30 June 1999 (30.06.1999)</b>  (30) Priority Data: <b>09/118,400 17 July 1998 (17.07.1998) US</b>  (60) Parent Application or Grant <b>SITARA NETWORKS, INC. [/]; O. YAO, Jie [/]; O. GOETZ, Thomas [/]; O. YAO, Jie [/]; O. GOETZ, Thomas [/]; O. PRAHL, Eric, L. ; O.</b>		Published
(54) Title: CONGESTION CONTROL (54) Titre: PROCEDE DE REGULATION DE L'ENCOMBREMENT  (57) Abstract  A new approach to congestion control includes features which overcome many of the limitations of the current congestion control approaches. The new approach uses a rate-based congestion control mechanism which uses a combination of multiple indicators of congestion (520, 522, 524). The transmission rate is decreased when there is an indication of congestion and the rate is increased when there is an indication of little or no congestion. The approach can also limit the transmission rate of multiple data streams destined to the same network node.		
(57) Abrégé  Une nouvelle approche de la régulation de l'encombrement comprend des caractéristiques qui surmontent plusieurs limitations des approches actuelles de régulation de l'encombrement. La nouvelle approche utilise un mécanisme de régulation de l'encombrement fondé sur le débit qui emploie une combinaison de plusieurs indicateurs d'encombrement (520, 522, 524). Le débit de transmission est réduit lorsqu'il existe une indication d'encombrement et le débit est accru lorsqu'il existe une indication de faible encombrement ou d'encombrement nul. L'approche peut également limiter le débit de transmission de plusieurs trains de données destinés au même noeud de réseau.		

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION  
International Bureau



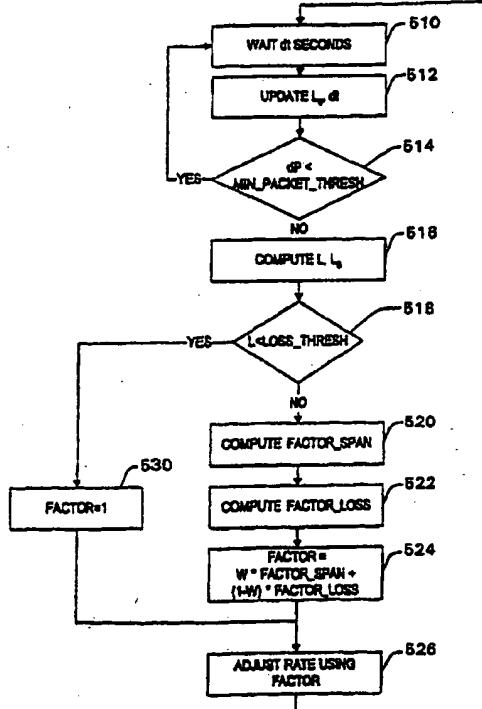
INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :  H04J 3/26		A1	(11) International Publication Number: WO 00/04665  (43) International Publication Date: 27 January 2000 (27.01.00)
<p>(21) International Application Number: PCT/US99/15071</p> <p>(22) International Filing Date: 30 June 1999 (30.06.99)</p> <p>(30) Priority Data: 09/118,400 17 July 1998 (17.07.98) US</p> <p>(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Application US 09/118,400 (CON) Filed on 17 July 1998 (17.07.98)</p> <p>(71) Applicant (for all designated States except US): SITARA NETWORKS, INC. [US/US]; Suite 3, 60 Hickory Drive, Waltham, MA 02154 (US).</p> <p>(72) Inventors; and (75) Inventors/Applicants (for US only): YAO, Jie [CN/US]; 44 Hanson Road, Newton, MA 02159 (US). GOETZ, Thomas [US/US]; 102 West Street, Foxboro, MA 02035 (US).</p> <p>(74) Agent: PRAHL, Eric, L; Fish &amp; Richardson, P.C., 225 Franklin Street, Boston, MA 02110-2804 (US).</p>		<p>(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, HR, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).</p> <p>Published <i>With international search report.</i></p>	

(54) Title: CONGESTION CONTROL

(57) Abstract

A new approach to congestion control includes features which overcome many of the limitations of the current congestion control approaches. The new approach uses a rate-based congestion control mechanism which uses a combination of multiple indicators of congestion (520, 522, 524). The transmission rate is decreased when there is an indication of congestion and the rate is increased when there is an indication of little or no congestion. The approach can also limit the transmission rate of multiple data streams destined to the same network node.



**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakhstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LX	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

**Description**

**5**

**10**

**15**

**20**

**25**

**30**

**35**

**40**

**45**

**50**

**55**

5

- 1 -

10

CONGESTION CONTROLBackground of the Invention

15 5 This invention relates to network communication protocols, such as communication protocols used in the Internet.

20 10 Internet communication is based on a layered model of communication protocols consistent with that published by the International Standards Organization (ISO). The 15 set of ISO protocol layers, or protocol stack, is numbered from one, at the lowest layer, to seven, at the application layer.

25 15 Communication over the Internet is based on packet-switching techniques. Addressing and transport of individual packets within the Internet is handled by the Internet Protocol (IP) corresponding to layer three, the 20 "network" layer, of the ISO protocol stack. This layer provides a means for sending data packets from one host to another based on a uniform addressing plan where individual computers have unique network addresses. By 25 making use of the IP layer, a sending computer is relieved of the task of finding a route to the destination host. However, packets may be lost or damaged due to random errors on data links or as a result of congestion within the network. Also, a sending host 30 may be able to provide data packets at a higher rate than can be accepted by a destination host, or than can be accepted by intermediate nodes or links of the network, thereby contributing to congestion within the network. The sending host is generally responsible for limiting 35 its rate of transmissions to avoid congestion in the network. This limiting of transmissions is implemented in software layered above the network layer.

40 45 35 At the next layer of the ISO protocol stack above the network layer, a transport layer (layer four)

55

- 2 -

10 protocol provides end-to-end communication between applications executing on different computers and regulates the flow of information between those applications. Rate control and error control are two 15 examples of regulations of the flow of information. Rate control addresses the rate at which data is transmitted into the network. In particular, rate control is one approach to congestion control. Error control addresses reliable delivery, for instance, providing error-free and 20 in-sequence delivery of data packets.

25 Today, the Transmission Control Protocol (TCP) is used almost exclusively to provide end-to-end reliable (i.e., error free) data streams between computers over the Internet. In the Internet, TCP is layered on the IP 30 15 network layer protocol. Software supporting use of the TCP protocol is provided on most popular operating systems, such as Microsoft Windows 95 and Windows NT, and most variants of Unix. An application using TCP is relieved of the details of creating or maintaining a 20 reliable stream to a remote application and simply requests that a TCP-based stream be established between itself and a specified remote system.

35 The success of TCP during last 20 years is due, at least in part, to its stable end-to-end congestion 40 25 control mechanism. TCP uses a window-based (or equivalently a credit-based) congestion control mechanism on each connection. For each connection, TCP limits the 45 30 number of bytes than can be sent that have not been acknowledged. In general, TCP implementations send as much data as possible, as soon as possible, without 45 35 exceeding the congestion window. TCP then waits for an acknowledgment of data in the window, or expiration of a timeout period, before it sends more data. The TCP congestion control mechanism adapts to network conditions 50 35 by dynamically modifying the size of the congestion

5

- 3 -

10

window. In general, the window is reduced quickly when packets are not delivered successfully. The window is increased slowly up to a maximum during periods when data is successfully delivered.

15

5

Summary

20

In a general aspect, this invention provides a new approach to congestion control. This new approach includes features which overcome many of the limitations of the current congestion control approaches. For instance, the new approach uses a rate-based congestion control mechanism which uses a combination of multiple indicators of congestion. The transmission rate is decreased when there is an indication of congestion and the rate is increased up to a predetermined maximum rate when there is an indication of little or no congestion. The approach can also limit the transmission rate of multiple data streams destined to the same network node.

25

30

In one aspect, in general, the invention is a method for congestion control in a data communication network by controlling a transmission rate a source of data transmits data onto a data network. The method features deriving multiple statistics from data communication from the source to a destination, the statistics providing indications of congestion on the data network. For instance, the statistics can include indicators of congestion such as a rate and a pattern of packet loss. The method also features adjusting the transmission rate to the destination in response to a combination of the derived statistics.

35

40

45

The method can also feature forming a group of data streams for transmission from the source, transmitting data from the group of data streams, and accepting acknowledgments of receipt of the transmitted data. As part of deriving the statistics related to

50

55

5

- 4 -

10 delivery of the data transmitted from the source, the transmissions of the data and the acknowledgments of receipt of the data can be monitored. The group of data streams can be formed so that they have a common  
15 5 destination on the data network, for example having a common host address on the Internet.

The method can include one or more of the following features.

20 The method can include computing a maximum  
10 transmission rate as a function of the multiple statistics and then limiting the transmission rate to the computed maximum transmission rate.

25 The statistics can include a rate of data loss and a pattern of data loss. In addition, the pattern of data  
15 loss can include lengths of sequences of lost data.

Adjusting the transmission rate can be performed in each of a sequence of time intervals.

30 In another aspect, in general, the invention is software stored on a computer readable medium. The  
20 software is for causing a computer to perform functions featuring deriving multiple statistics from data communication from a source of data over a data network to a destination. The statistics provide indications of congestion of the data network. The functions also  
25 feature adjusting a transmission rate from the source to the destination in response to a combination of the derived statistics.

40 In another aspect, in general, the invention is a congestion control apparatus. The apparatus features a  
30 rate updater for determining a maximum rate of data transmission to a destination over a data network. The rate updater determines the maximum rate using a combination of a plurality of statistics derived from communication with the destination. The apparatus also  
45 35 features storage associating the destination with

50

55

5

- 5 -

10 determined maximum rate and a transmission throttle from limiting a rate of data transmission to the destination based on the stored maximum rate.

15 5 Aspects of the invention include one or more of the following advantages.

20 10 Use of a congestion control mechanism which is separate from error control mechanisms allows maintaining high throughput for applications which can tolerate a modest error rate.

25 15 The rates of a group of connections to a common destination can be controlled together. Patterns of packet loss are monitored on the group of streams, thereby providing improved indicators of congestion compared to indicators based solely on the individual data streams.

30 30 Also, by not assuming that all packet loss is due to congestions, the invention can provide high throughput networks with relatively high random data loss (e.g., greater than 1% loss), such as is typical on some wireless data networks. Furthermore, data sent according to this invention can be less bursty than using other congestion control approaches, thereby improving overall network performance.

35 35 25 Other features and advantages of the invention will be apparent from the following description, and from the claims.

#### Description of the Drawing

40 45 FIG. 1 shows two network nodes coupled through a data network;

45 50 FIG. 2 illustrates a sequence of packets with multiple spans of packet loss;

50 FIG. 3 shows ranges of two statistics used to compute transmission rate changes;

55

5

- 6 -

10

FIG. 4 is a flowchart of a connection procedure;

15

FIG. 5 is a flowchart of a rate adjustment procedure;

FIG. 6 shows software elements of a rate controller; and

FIG. 7 shows hardware elements of a network node.

Description

20

Referring to FIG. 1, two network nodes (i.e., general or special purpose computers) 110A and 110B are

10 coupled through a data network 100. Communication passing between the network nodes, in general, passes over multiple links in data network 100. For instance, in the exemplary data network shown in FIG. 1, communication passing from network node 110A to network

15 node 110B passes over links 102, 104, and 106.

25

Congestion in data network 100 can occur for a variety of reasons. For instance, congestion can occur at intermediate points in the network. In this example, link 104 has relatively lower capacity than links 102 and

20 106, or must share a comparable capacity with data arriving from other links. Therefore, if data passes

30 over link 102 at the full rate supported by that link, the data must be queued at intermediate point 103 before passing over link 104 at a lower rate. Because the queue at point 103 has a bounded capacity, if network node 110A continues to send at a high rate, some of that data will

35 eventually be lost at point 103 when its queue overfills. When data is lost in this way, in general, a series of data packets sent from network node 110A will be lost.

40

45 30 In each network node 110A, 110B software modules include one or more applications 112 each of which can establish multiple data streams with other applications through a transport layer 114. Transport layer 114 in turn communicates with a network layer 118 to support

50

55

5

- 7 -

10 communication between applications on different network nodes. Each network layer communicates with a corresponding network interface controller 120 which provides a physical connection to data network 100.

15 5 Using these components, an application 112 on network node 110A can communicate with an application 112 on network node 110B.

20 Transport layer 114 include a rate controller 116. Rate controller is used to limit the rate that packets 10 are sent over a connection between applications on different network nodes. Rate controller 116 on a network node separately limits the total data transmission rate of all data streams from applications 25 on that network node to each remote network node. In the 15 situation described above in which data arrives over link 102 at a rate higher than can be accepted by link 104 and data is dropped, rate controller 116 at node 110A is 30 designed with the goal of reducing the rate that data is sent over link 104 thereby relieving the congestion at 20 point 103.

Congestion Indicators

35 Rate controller 116 adapts the transmission rate based on multiple indicators of congestion. Not only is an average rate of packet loss used, but the pattern of 25 those losses is also used. Referring to FIG. 2, a sequence of packets 200 sent by one node to another is 40 illustrated. The sequence of  $dP=17$  packets includes successfully received packets 210, illustrated as solid squares, and  $DL=6$  lost or damaged packets 212, 45 illustrated as broken squares. The lost or damaged packets occur in  $ds=3$  "loss spans," each of which is a consecutive subsequence of lost packets. Rate controller 116 computes two statistics for such a sequence of sent 50 packets 200. The first is a loss rate,  $L$ , which is the fraction of packets that are lost in the sequence. In

5

- 8 -

10 this exemplary sequence,  $L=6/17=0.35$ . The second statistic relates to the pattern of loss. Rate controller 116 computes a "cluster loss ratio,"  $L_s$ , defined as the ratio of the number of loss spans to the 15 number of lost packets. In this exemplary sequence,  $L_s=3/6=0.50$ . Note that  $L_s$  is close to 1 if the pattern of packet loss is "random" consisting of isolated lost packets. On the other hand,  $L_s$  is small if the pattern of loss consists of long subsequences of consecutive lost 20 packets. Long subsequences of lost packets are an indication of congestion in the network. For instance, an overfull buffer at an intermediate node in the network will not accept new data until it has cleared its backlog. Therefore, in general, multiple sequential 25 15 packets arriving at that intermediate node will be lost.

30 Rate controller 116 also uses a longer-term statistic of packet loss. Specifically, an average rate of packet loss,  $L_o$ , is tracked. Packet loss in a particular sequence is expected to be close to  $L_o$  if the 20 loss is due to random errors, such as errors on a data link. Rate controller 116 uses the amount by which the packet loss rate differs the average loss rate as an 35 indication of congestion or lack of congestion.

Rate Adjustment

25 Rate controller 116 repeatedly adjusts the packet transmission rate,  $R$  (packets per second), based on the 40 sequence of packets sent since the last adjustment of rate. Based on the rate and pattern of packet loss, rate controller 116 either increases  $R$ , decreases  $R$ , or leaves 30  $R$  unchanged.

45 Referring to FIG. 3, rate controller computes an excess loss rate,  $L-L_o$ , and a loss ratio,  $1-L_s$ , in order to adjust the transmission rate. These two quantities are illustrated in a two-dimensional plane with axes 310 35 and 320. Note that  $L-L_o$  can range from -1.0 to 1.0 while

50

55

5

- 9 -

10        $1-L_s$  can ranges from 0.0 to 1.0. When  $L-L_0$  is close to 1.0, the loss rate is high relative to a low average loss rate. When  $1-L_s$  is close to 1.0, lost packets occur in relatively long spans, indicating congestion. When  $L-L_0$  15 is close to -1.0, the loss rate is low relative to a high average loss rate. When  $1-L_s$  is close to 0.0, lost packets occur in relatively short spans. In general, when the loss rate is high and the loss spans are long (i.e., the top right region of the graph), rate 20 controller 116 decreased the transmission rate. When the loss rate is low and the spans are short (i.e., the lower left region of the graph), rate is generally increased.

25       Two ranges are defined for each variable. On the excess loss rate axis 310, a loss hysteresis threshold 15 (LOSS\_HYST) 312 defines a range 314 between LOSS\_HYST and 1.0. In this range, an excess loss rate contributes to a decrease in transmission rate. The negative of the loss hysteresis threshold (-LOSS\_HYST) 316 defines a range 318 30 from -LOSS\_HYST to -1.0 in which the excess loss rate 20 contributes to an increase in transmission rate.

35       On loss ratio axis 320, an upper span loss ratio threshold (UPPER\_SPAN\_THRESH) 326 defines a range 328 between UPPER\_SPAN\_THRESH and 1.0 in which a loss ratio 40 contributes to a decrease in transmission rate. A lower span loss ratio threshold (LOWER\_SPAN\_THRESH) 322 defines a range 324 between 0.0 and LOWER\_SPAN\_THRESH in which a loss ratio contributes to an increase in transmission rate.

45       A value of 0.06 for HYST\_THRESH, and values of 30 0.09 and 0.286 for LOWER\_SPAN\_THRESH and 45 UPPER\_SPAN\_THRESH, respectively, have been used successfully.

50       In some ranges of values of the two variables, for example, when the excess loss rate is greater than 35 HYST\_THRESH (i.e., in range 314) and the loss ratio is

55

5

- 10 -

10 less than LOWER\_SPAN\_THRESH (i.e., in range 316) the  
excess loss rate and the loss ratio contribute to  
decreasing and increasing the rate, respectively. The  
relative contributions of these two factors determine  
15 5 whether the transmission rate is in fact increased or  
decreased. Similarly, when the excess loss rate is less  
than -HYST\_THRESH (i.e. in range 318) and the loss ratio  
is greater than UPPER\_SPAN\_THRESH (i.e., in range 328),  
the two factors also compete to determine whether the  
20 10 transmission rate actually increases or decreases.

Based on the loss ratio and excess loss rate of a  
sequence of packets, rate controller 116 computes two  
25 factors, a span factor (FACTOR\_SPAN) and a loss factor  
(FACTOR\_LOSS). These factors are in a range -1.0 to 1.0.  
15 If the loss ratio ( $1-L_g$ ) exceeds the upper span loss ratio  
threshold, UPPER\_SPAN\_THRESH, the span factor is a  
normalized amount by which it exceeds the threshold. In  
particular, the span factor is computed as  
30

$$20 \text{ FACTOR\_SPAN} = \frac{((1-L_g) - \text{UPPER\_SPAN\_THRESH})}{(1.0 - \text{UPPER\_SPAN\_THRESH})}$$

35 If the loss ratio is less than the lower span loss ratio  
threshold, then the span factor is computed as

$$40 \text{ FACTOR\_SPAN} = \frac{-(1-L_g)}{\text{LOWER\_SPAN\_THRESH}}$$

25 Note that in the first case, the computed span factor is  
in the range 0 to 1.0 while in the second case, the  
computed span factor is in the range -1.0 to 0.

45 45 Rate controller 116 computes the loss factor in a  
similar manner. In particular, if the excess loss rate  
exceeds the loss hysteresis threshold, then the loss  
30 factor is computed as

50

55

5

- 11 -

10

$$\text{FACTOR\_LOSS} = ((L-L_0)-\text{HYST\_THRESH}) / (1-\text{HYST\_THRESH})$$

15

Similarly, if the excess loss rate is lower than the negative loss hysteresis threshold, the loss factor is computed as

20

$$5 \quad \text{FACTOR\_LOSS} = ((L-L_0)+\text{HYST\_THRESH}) / (1-\text{HYST\_THRESH})$$

25

Note that in the first case, the computed loss factor is in the range 0 to 1.0 while in the second case it is in the range -1.0 to 0.

30

To illustrate this calculation, consider a pair of values illustrated by the point 336. FACTOR\_SPAN is negative with a magnitude equal to the ratio of the length of line segment 332 to the length of range 328, and FACTOR\_LOSS is negative with a magnitude equal to the ratio of the length of line segment 334 to the length of range 314.

35

Having computed FACTOR\_LOSS and FACTOR\_SPAN, rate controller 116 computes a weighted average of the these factors to derive a combined factor. The relative weighting of the factors is configurable, according to a span ratio weight, W, which is in the range 0.0 to 1.0. The combined factor is computed as

40

$$\text{FACTOR} = W * \text{FACTOR\_SPAN} + (1-W) * \text{FACTOR\_LOSS}$$

45

A value of W=0.67 for the span ratio weight has been used successfully.

25 If the combined factor is positive, then the rate is increased. If the factor is negative, the rate is decreased. Specifically, if FACTOR>0 and the current rate is R\_OLD, then the new rate, R\_NEW, is computed as

50

$$R_{\text{NEW}} = (1 + \text{FACTOR/CHANGE\_FACTOR\_UP}) * R_{\text{OLD}}$$

55

5

- 12 -

10 If FACTOR&lt;0, then R\_NEW is computed as

$$R_{\text{NEW}} = (1 + \text{FACTOR}/\text{CHANGE\_FACTOR\_DOWN}) * R_{\text{OLD}}$$

15 The values of approximately 2.0 and 1.75 for the  
CHANGE\_FACTOR\_UP and CHANGE\_FACTOR\_DOWN, respectively,  
5 have been used successfully. These values determine time  
constants of rate increases or decreases. Using these  
values, R\_NEW is within the approximate range of 0.4 to  
20 1.5 times R\_OLD.

After computing R\_NEW according to the formulas  
10 above, R\_NEW is limited to be within a predetermined  
range from a minimum rate to a maximum rate. The minimum  
25 rate is a configurable constant rate. A value of 500  
bytes/second can be used. The maximum rate is set based  
on the maximum rate that is negotiated when connections  
15 are established between the local and the destination  
node.

30 The above procedure is only applied if the loss  
rate, L, for a sequence of packets, is above a loss  
threshold, LOSS\_THRESH. If L<LOSS\_THRESH, then the rate  
20 is increased according to

35

$$R_{\text{NEW}} = (1 + 1/\text{CHANGE\_FACTOR\_UP}) * R_{\text{OLD}}$$

40 and limited by the maximum predetermined rate. A value  
of 0.06 for LOSS\_THRESH has been used successfully. In  
this way, the rate increases up to the maximum while the  
25 absolute loss rate is low.

#### Adjustment Periods

45 This rate updating procedure described above is  
applied to successive sequences of sent packets.  
Periodically, every dt seconds, a rate adjustment is  
30 considered by rate controller 116. The update time, dt,  
is adapted to each destination and kept at a value of

50

55

5

- 13 -

10 approximately 6 times the round-trip time of  
communication to the destination and back. Since the  
rate adjustment relies on estimates of the loss rate and  
the loss ratio, if fewer than a minimum number of  
15 packets, MIN\_PACKET\_THRESH, have been sent since the rate  
adjustment, the rate adjustment is deferred for another  
dt seconds. A value of 8 for MIN\_PACKET\_THRESH has been  
used successfully.

20 After each dt seconds, rate controller 116 updates  
10 its average loss rate,  $L_b$ , to be the ratio of the number  
of packets that were successfully received to the number  
of packets that were sent. Alternate averaging  
approaches, such as a decaying average can be used. Rate  
25 controller 116 also updates its estimate of the round-  
15 trip time to the destination.

30 Note that the above technique relies on the  
receiving node sending selective acknowledgments of  
packets to the sending node. Referring back to FIG. 2,  
after packets 3 and 4 are lost, the receiving network  
20 node receives packet 5. The receiving node acknowledges  
receipt of packet 5. This acknowledgment allows the  
sending node to determine that packets 3 and 4 have been  
35 lost. At the end of every dt seconds interval, the  
controller 116 only considers packets up to the most  
25 recently acknowledged packet. Therefore, packets that  
are still "in flight" are not considered.

40 Connection and Rate Adjustment Procedures

The connection procedure and subsequence rate  
adjustment is summarized in the flowcharts shown in FIGS.  
30 4 and 5. Referring to FIG. 4, transport layer 114 (FIG.  
45 1) receives a request to establish a data path with  
destination network node (step 410). The transport layer  
exchanges connection information with the destination  
node (step 412). Included in that information is the  
35 maximum data transmission and receiving rates supported

50

55

5

- 14 -

10 by each of the network nodes. If there is no other  
connection to the destination node (step 414), a new  
destination rate manager is created (step 416). As is  
described more fully below, the destination rate manager  
15 5 contains information needed to control the transmission  
rate to a particular destination. If other connections  
are active to this destination, the connection is linked  
to an existing destination rate manager (step 418). The  
transmission rate to the destination node is then  
20 10 controlled by the transport layer using the destination  
rate manager for that destination (step 420).

The rate adjustment procedure for a particular  
destination is summarized in the flowchart shown in FIG.  
25 5. After a communication session is set up, rate  
15 controller 116 (FIG. 1) waits for the expiration of a dt  
duration interval (step 510). The rate controller  
updates the long term packet loss rate,  $L_o$ , using the most  
30 recent sequence of sent packets, and updates the rate  
update time, dt, based on the round-trip time (step 512).  
20 20 If the number of packets sent since the last rate update  
is less than a threshold (step 514) the controller  
returns to wait for the expiration of another interval  
35 (step 510). Otherwise, based on the sequence of sent  
packets since the last rate update, the rate controller  
25 25 computes the loss rate,  $L$ , and the loss ratio,  $L_s$  (step  
516). If the loss rate is not less than a threshold  
40 (step 518), the controller computes FACTOR\_SPAN (step  
520) and FACTOR\_LOSS (step 522) according to the formulas  
presented above, and then combines these to compute the  
30 30 overall FACTOR (step 524). If, on the other hand, the  
loss rate is less than the threshold (step 518) FACTOR is  
45 set to 1. Based on the computed FACTOR, the controller  
then adjusts the transmission rate (step 526) according  
to the formulas presented above. The rate controller  
35 50 then returns to wait for the end of another dt interval

55

5

- 15 -

10 (step 510).

Transport Layer Modules

15 Referring to FIG. 6, the controller 116 includes several modules. Transport layer 114 supports 5 connections from multiple applications 112. Each application can concurrently have open connections to multiple destinations. Communication to and from each destination passes through rate controller 116 in transport layer 114.

20 Rate controller 116 is implemented using a destination mapper 614 through which all connections pass, and a single destination rate controller 612 for each destination with which any application 112 is 25 communicating. Destination rate controllers 612 are 15 created when an initial connection to a new destination is established (FIG. 4, step 418). Subsequent connections to the same destination on behalf of any application 112 use the same destination rate controller 612 (FIG. 4, step 418). Once all connections to a 30 destination are closed, the destination rate controller 20 for that destination is "destroyed." Destination rate controllers are implemented as C++ objects.

35 When an application 112 sends a packet of data to a destination, that packet passes from the application to 25 destination mapper 614. Based on the destination, destination mapper 614 passes the data to a particular destination rate controller 612.

40 Within each destination rate controller 612, a transmission throttle 620 limits the rate of data 30 transmission to the destination. Transmission throttle 620 is implemented by periodically (e.g., every 200 45 milliseconds) determining how much pending data for each destination can be sent to network layer 118 without exceeding the calculated transmission rate for that 35 destination. Data that cannot be sent is buffered by

50

55

5

- 16 -

10 transmission throttle 620. In each interval, transmission throttle 620 increments a "credit" based on the duration of the interval and the transmission rate and decrements the credit based on the amount of data  
15 sent. The amount of data sent is limited to keep the credit non-negative. The credit is bounded to not grow beyond a specific amount, in particular, it is bounded by the transmission rate times the duration of two update intervals.

20 10 In the return direction, data from remote nodes pass from network layer 118 to destination mapper 614 and then to the destination applications 112.

25 15 Each destination rate controller 116 includes a table 624 that includes information needed to control the rate of that destination. In particular, table 624 includes the current maximum transmission rate (R), the current estimate of average loss rate ( $L_e$ ), the number of packets sent since the last rate update (dP), the number of packets lost since the last rate update (dL), and the 30 20 number of spans of lost packets since the last rate update (dS). Transmission throttle 620 limits the number of packets so as not to exceed the current maximum transmission rate (R). Destination rate controller 612 also includes a rate updater 622 which monitors the 35 25 packet transmissions and acknowledgments to and from its corresponding destination, and updates table 624 based on the rate and pattern of lost packets.

40 Alternative software architectures of rate controller 116 can also be used. For instance, a single 30 45 transmission throttle module and a single rate updater module can be used for all connections. Instead of creating separate destination rate controller objects, one for each destination, each with a separate table 624 holding information related to the rate control for that 35 destination, a common table can be used associating each 50

5

- 17 -

10 destination with information related to the rate control for that destination. The single transmission throttle and rate updater use and update appropriate records in the common table based on the destination of  
15 communication.

15 Referring to FIG. 7, a network node implements the software modules shown in FIG. 6. The network node includes a processor 712 and working memory 710. Working memory holds rate table 620 (FIG. 6) as well as the code  
20 10 that implements transmission throttle 610 and rate updater 612, as well as other software modules. Network node also includes permanent program storage 714 and network interface controller 120 which coupled the network node to data network 100.

25 15 In the above embodiment, transmission rate is controlled separately for each destination node. Alternatively, transmission rate can be controlled for other groupings of connections and congestion statistics computed for those groups. For example, individual  
30 20 connections can be individually controlled, or groups of connections that share particular characteristics can be controlled together.

35 25 Although not shown, transport layer 114 can include other modules that serve functions that are well known to one skilled in the art. In particular, transport layer 114 can include an error control module that provides a reliable data stream to application 112, and a flow control module to limit the amount of unacknowledged data that is sent on each individual  
40 30 connection.

45 35 Other embodiments can use alternative indicators of congestion or other ways of combining the loss rate and loss ratio indicators. For instance, quantized span and loss factors can be computed rather than computing the floating point versions described above. Also,

50

55

5

- 18 -

10 rather than setting specific thresholds for the indicator variables, other functions mapping the indicator variables and a current rate to a new rate can be used.

15 It is to be understood that the foregoing description is intended to illustrate and not limit the scope of the invention, which is defined by the scope of the appended claims. Other aspects and modifications are within the scope of the following claims.

20

What is claimed is:

25

30

35

40

45

50

55

**Claims**

**5**

**10**

**15**

**20**

**25**

**30**

**35**

**HIS PAGE BLANK (USPTO)**

**40**

**45**

**50**

**55**

10 1. A method for congestion control in a data communication network by controlling a transmission rate at which a source of data transmits data onto a data network comprising:

15 5 deriving a plurality of statistics related to delivery of data transmitted from the source, the statistics providing indications of congestion on the data network;

20 10 adjusting the transmission rate from the source in response to a combination of the derived statistics.

25 2. The method of claim 1 further comprising: forming a group of data streams for transmission from the source;

30 15 transmitting data from the group of data streams;

35 20 accepting acknowledgments of receipt of the transmitted data from the group of data streams; and

30 25 wherein deriving the statistics related to delivery of the data transmitted from the source includes monitoring the transmissions of the data and monitoring the acknowledgments of receipt of the data.

35 30 3. The method of claim 2 wherein deriving the statistics further includes combining acknowledgements for different data streams in the group of data streams.

40 40 25 4. The method of claim 2 wherein forming a group of data streams includes forming a group of data streams which have a common destination on the data network.

45 45 30 5. The method of claim 1 further comprising: computing a maximum transmission rate as a function of the plurality of statistics; and wherein adjusting the transmission rate includes

5

- 20 -

10 limiting the transmission rate to the computed maximum transmission rate.

15 6. The method of claim 1 wherein deriving the plurality of statistics includes monitoring a rate of 5 data loss and a pattern of data loss.

20 7. The method of claim 6 wherein monitoring the pattern of data loss includes monitoring lengths of 25 sequences of lost data.

8. The method of claim 1 wherein deriving the 10 statistics and the adjusting of the transmission rate is 25 performed in each of a sequence of time intervals.

9. The method of claim 1 wherein adjusting the transmission rate includes:  
30 computing a first factor related to a rate of data 15 loss;  
computing a second factor related to lengths of data loss;  
35 combining the first and second factors;  
adjusting the transmission rate according to the 20 combined factor.

10. Software stored on a computer readable medium 40 for causing a computer to perform the functions of:  
deriving a plurality of statistics related to 25 delivery of data transmitted from a source of data over a data network, the statistics providing indications of 45 congestion on the data network;  
adjusting the transmission rate from the source in response to a combination of the derived statistics.

11. The software of claim 10 further causing the 50

55

5

- 21 -

10 computer to perform the functions of:  
forming a group of data streams for transmission  
from the source;  
transmitting data from the group of data streams;  
15 5 and  
accepting acknowledgments of receipt of the  
transmitted data; and  
wherein deriving the statistics related to  
delivery of the data transmitted from the source includes  
20 10 monitoring the transmissions of the data and monitoring  
the acknowledgments of receipt of the data.

12. A congestion control apparatus comprising:  
25 a rate updater for determining a maximum rate of  
data transmission to a destination over a data network,  
15 the rate updater determining the maximum rate using a  
combination of a plurality of statistics derived from  
30 communication with the destination;  
storage associating the destination with  
determined maximum rate; and  
20 a transmission throttle from limiting a rate of  
data transmission to the destination based on the stored  
35 maximum rate.

40

45

50

55

1/7

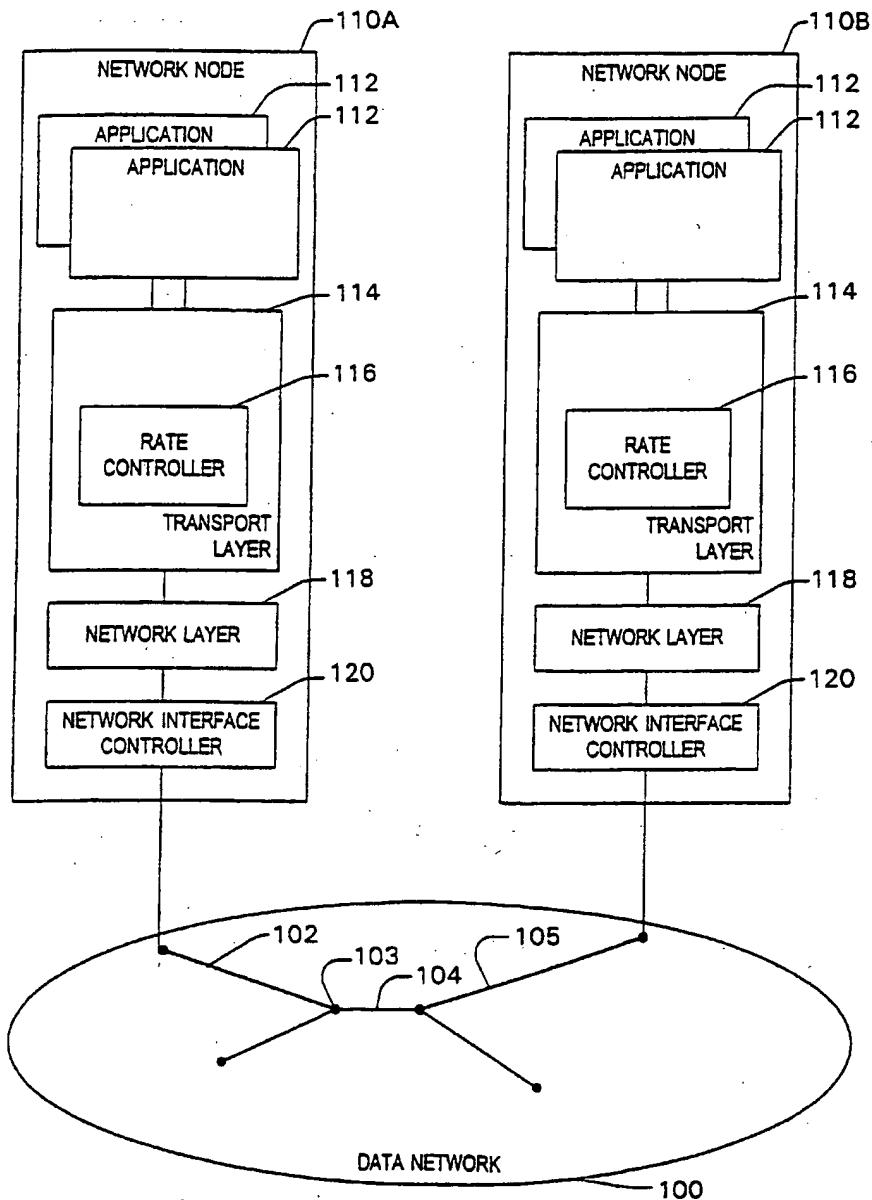


FIG. 1

SUBSTITUTE SHEET (RULE 26)

2/7

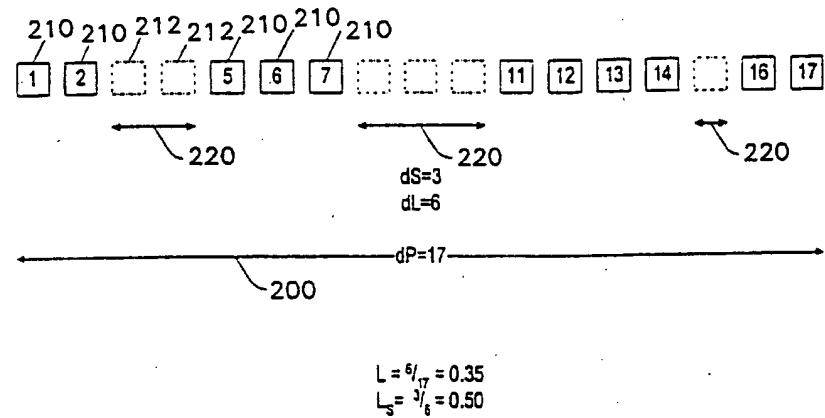


FIG. 2

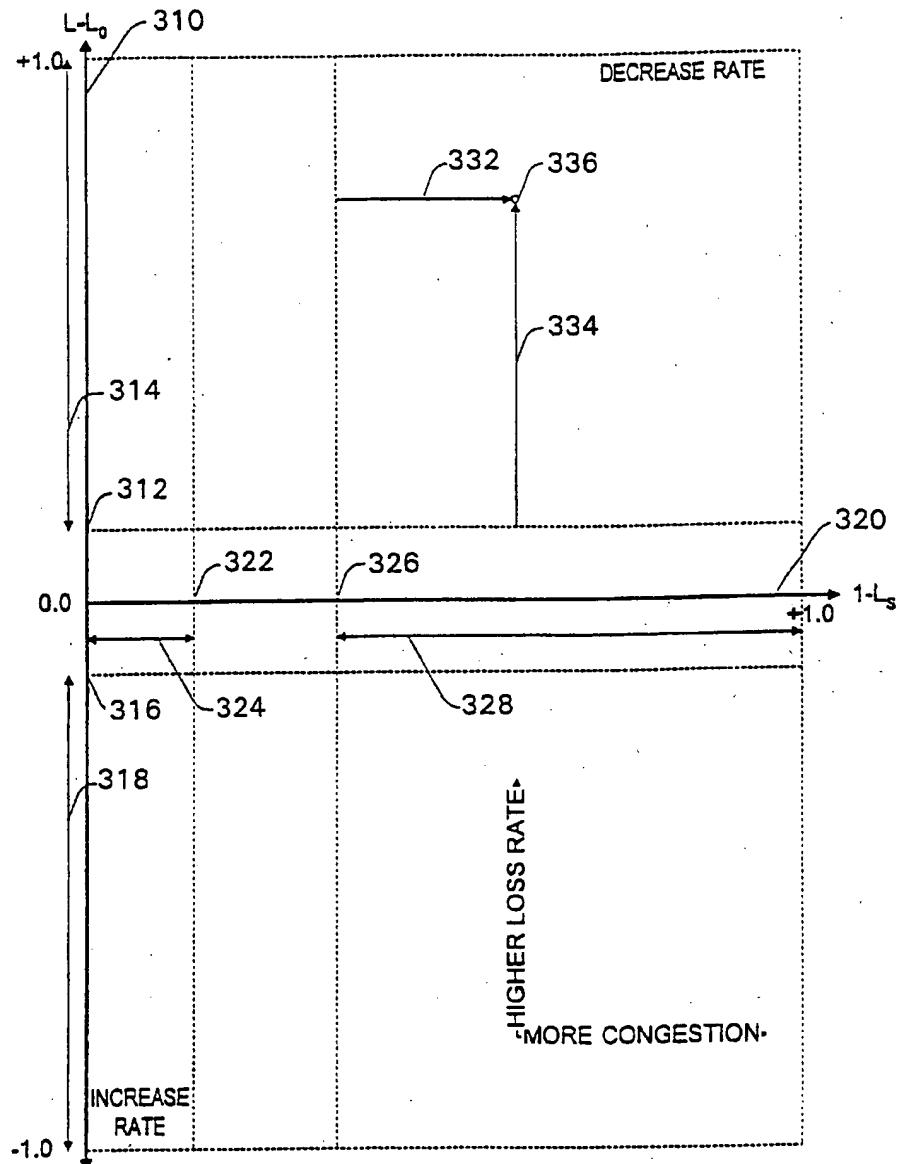


FIG. 3

4/7

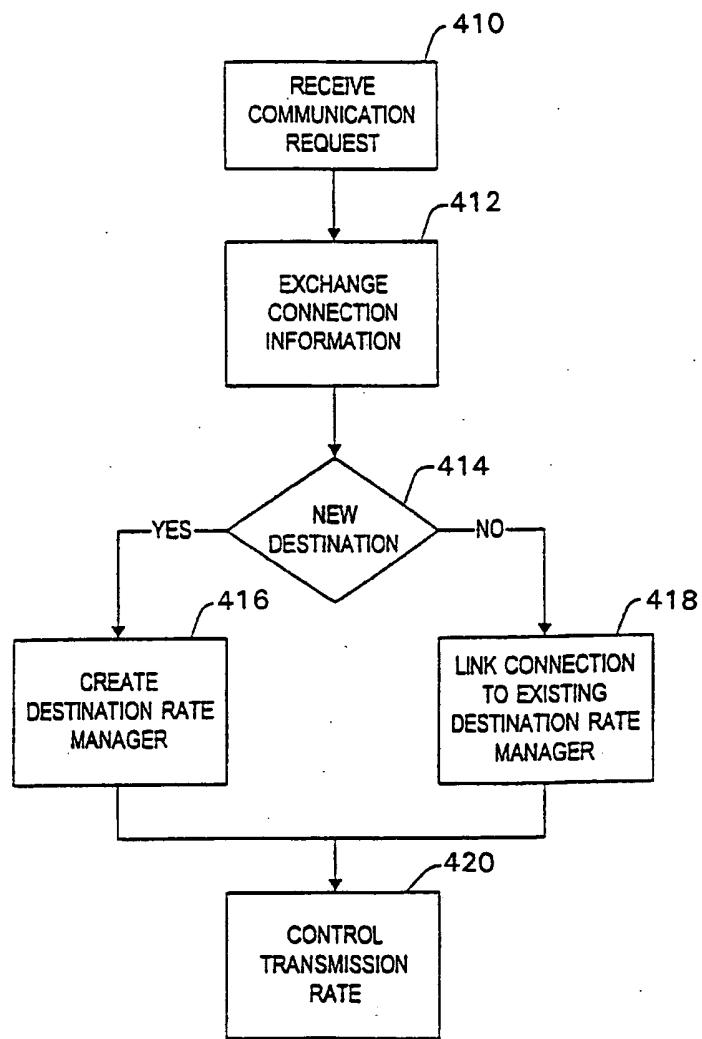


FIG. 4

5/7

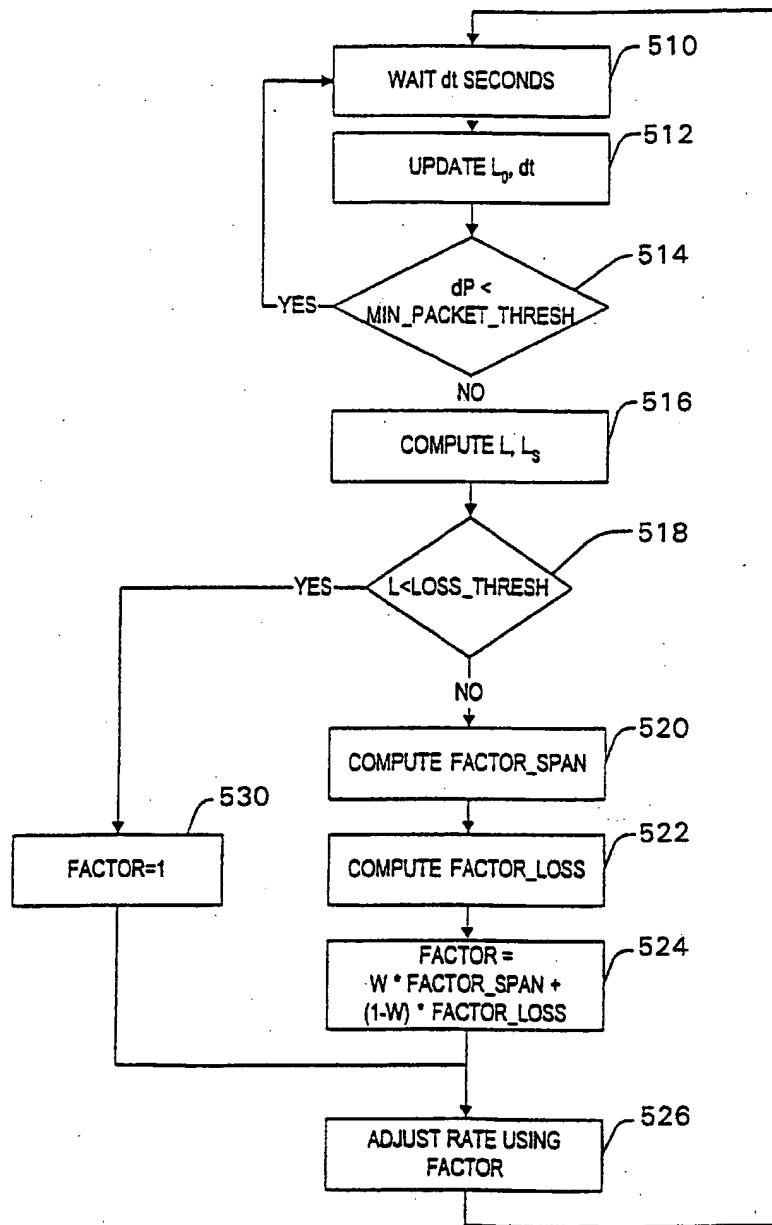


FIG. 5  
SUBSTITUTE SHEET (RULE 26)

6/7

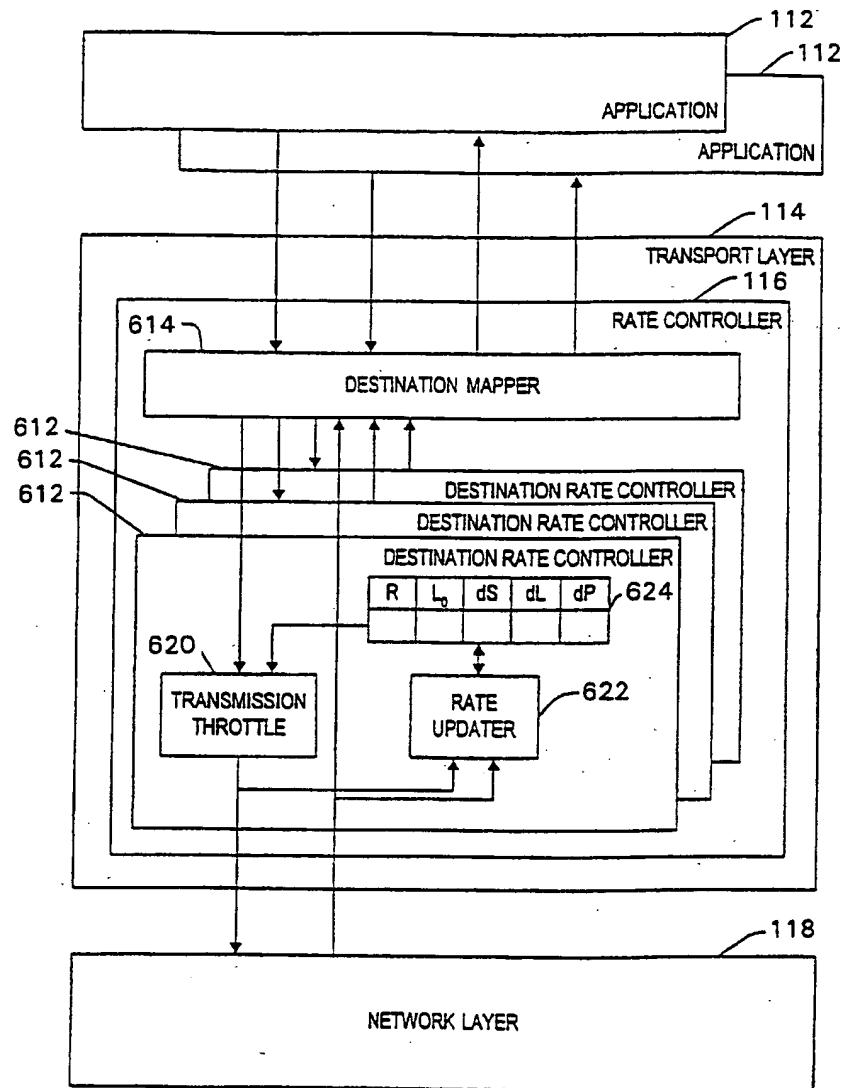
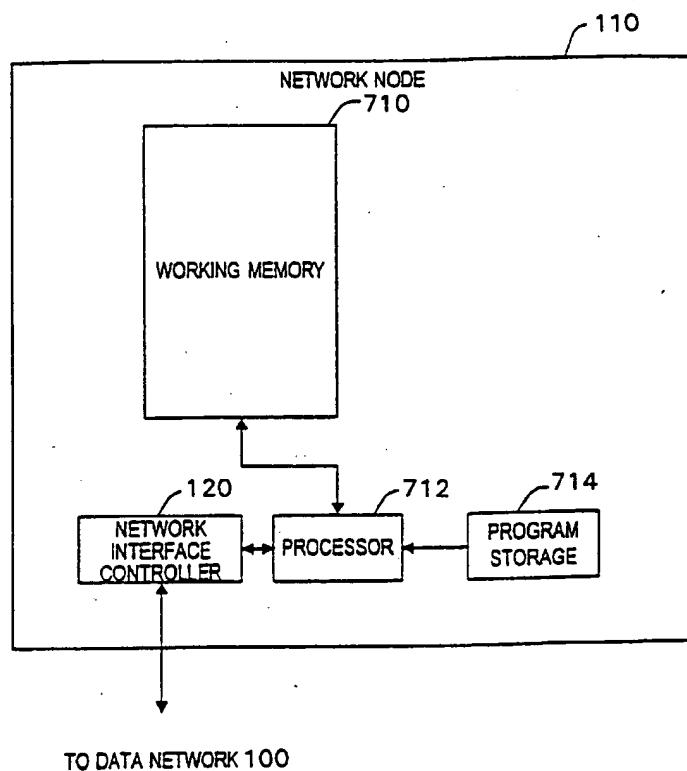


FIG. 6

7/7

**FIG. 7**

## INTERNATIONAL SEARCH REPORT

International application No.  
PCT/US99/15071

<b>A. CLASSIFICATION OF SUBJECT MATTER</b>		
IPC(6) :H04J 3/26 US CL :370/229, 230, 231, 232, 233, 234, 252, 253 According to International Patent Classification (IPC) or to both national classification and IPC		
<b>B. FIELDS SEARCHED</b>		
Minimum documentation searched (classification system followed by classification symbols)  U.S. : 370/229, 230, 231, 232, 233, 234, 252, 253		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)  APS, IEEE search terms: congestion, packet, lost, loss		
<b>C. DOCUMENTS CONSIDERED TO BE RELEVANT</b>		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5,650,993 A (LAKSHMAN et al) 22 July 1997, column 4, line 54 to column 5, line 20; column 5, line 41 to column 6, line 21; and column 6, lines 24-26.	1-5, 8 and 10-12
X	US 5,633,861 A (HANSON et al) 27 May 1997, see Abstract.	1
X	WILLIAMSON, C.L. "Dynamic Bandwidth Allocation Using Loss-Load Curves" IEEE/ACM TRANSACTIONS ON NETWORKING, Vol. 4, No. 6, December 1996, see Abstract.	1
<input type="checkbox"/> Further documents are listed in the continuation of Box C. <input type="checkbox"/> See patent family annex.		
<p>* Special categories of cited documents:</p> <p>"A" document defining the general state of the art which is not considered to be of particular relevance</p> <p>"B" earlier document published on or after the international filing date documents which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)</p> <p>"C" document referring to an oral disclosure, use, exhibition or other means</p> <p>"D" document published prior to the international filing date but later than the priority date claimed</p> <p>"E" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention</p> <p>"F" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone</p> <p>"G" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art</p> <p>"H" document member of the same patent family</p>		
Date of the actual completion of the international search  20 AUGUST 1999	Date of mailing of the international search report  26 OCT 1999	
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer  MELVIN MARCELO <i>F. M. Marce洛</i> Telephone No. (703) 305-4800	